Avant-propos du thème “ Le latin dans le texte ”.

Monique Goullet

Ceci est l'avant-propos du thème " Le latin dans le texte ", coordonné par M. Goullet et N. Bouloux, *Médiévales*, 42 (printemps 2002).

Avant-propos

Ce numéro consacré à la fois au latin et aux outils de statistique textuelle voudrait faire mentir une critique, à notre sens injuste, stigmatisant " l'effroi de la plupart de nos collègues devant le moindre chiffre ", " le peu d'attention accordé à l'examen détaillé et littéral des textes, l'attitude préréflexive qui consiste à croire que la plupart des textes sont transparents ", " l'attitude fondamentalement " positiviste ", c'est-à-dire l'incapacité à raisonner à propos des relations entre réalité, représentations et mots "[1]. Il est vrai que la tradition universitaire française laisse dépourvus de formation philologique et linguistique la plupart des historiens, et privés de formation historique la majorité des latinistes, les deux corporations étant par ailleurs aussi rarement l'une que l'autre instruites de la statistique. Il est exact aussi que les " littéraires " (entendons les historiens et les philologues en l'occurrence) éprouvent souvent une vive répulsion à l'égard des tableaux statistiques, et qu'ils agissent à cet égard avec le mépris du renard de La Fontaine en face des raisins qui lui échappent. S'ajoutent aujourd'hui les progrès énormes de l'informatique, avec l'arrivée sur le marché de quantité de logiciels de traitement lexical performants et peu onéreux, qui devraient mettre l'outil à la portée de tous, mais qui au lieu de cela creusent l'écart entre les initiés et les autres, faute d'une (in)formation adéquate.

Nous avons fait le pari que les historiens, à défaut de savoir le faire eux-mêmes, étaient persuadés de l'utilité de " compter " et de " traiter " les mots[2], et que les spécialistes étaient disposés à les y aider, avec toute leur compétence théorique et technique, mais aussi avec toute la distance critique acquise grâce à l'expérience de ces méthodes quantitatives. Nous avons donc invité autour d'une table, aimablement prêtée par Jacques Dalarun à l'IRHT, d'une part ces spécialistes des méthodes statistiques : Sylvie Mellet, latiniste et directrice du laboratoire " Bases, corpus et langage " de l'Université de Nice ; Michel Dubrocard (latiniste) et Xuan Luong (mathématicien), enseignants-chercheurs et membres de ce même laboratoire niçois ; Etienne Evrard, professeur émérite et chercheur au Lasla de Liège (Laboratoire d'analyse statistique des langues anciennes), d'autre part des philologues : Anita Guerreau-Jalabert et Bruno Bon, chercheurs au Novum Glossarium Latinitatis Medii Aevi (" Nouveau Du Cange ", Institut de France) ; Anne-Marie Turcan-Verkerk, chercheur à l'IRHT[3], et des historiens médiévistes : Michel Parisse, professeur à Paris I ; Monique Paulmier-Foucart et Marie-José Gasse (Atelier de Recherche sur les textes médiévaux et leur traitement assisté, Artem, Nancy)[4] et Nicholas Brousseau, doctorant à Paris I. Les lignes qui suivent ont pour objet de présenter les principales problématiques abordées, et de faciliter la lecture des articles par quelques

[1] - Alain Guerreau, " Pourquoi (et comment) l'historien doit-il compter les mots ? ", *Histoire et Mesure*, IV-1/2 (1989), p. 81-105, spéc. p. 81.

[2] - Beaucoup d'entre eux l'ont déjà fait, dont on trouvera les titres dans les notes des différents articles et dans la bibliographie finale. Pour ce qui est de l'utilisation de l'outil informatique, il faut mentionner ici *Informatique et Histoire médiévale*. Communications et débats de la Table ronde du CNRS organisée par l'Ecole Française de Rome et l'Institut d'Histoire médiévale de l'Université de Pise, Rome, 20-22 mai 1975, présentés par L. Fossier, A. Vauchez, C. Violante, Ecole Française de Rome, 1977, travail pionnier du temps des cartes perforées, évidemment tout à fait dépassé aujourd'hui, mais qui a joué un rôle moteur en son temps ; la question philologique n'y est cependant pas abordée. Un second jalon a été posé par *La lexicographie du latin médiéval et ses rapports avec les recherches actuelles sur la civilisation du moyen âge*, éd. Yves Lefèvre, Paris, 1981 (Colloque international du CNRS, octobre 1978), dont les deux derniers chapitres sont consacrés aux méthodes de lexicographie assistées par l'informatique, et aux grands dictionnaires en cours de réalisation.

[3] - A.- M. Turcan-Verkerk a eu la gentillesse de venir exposer ses recherches concernant le traitement statistique de la prose rythmée ; celles-ci ont déjà fait l'objet d'un précédent article, auquel nous renvoyons : A.- M. Turcan-Verkerk et Ph. Verkerk, " Un programme informatique pour l'étude de la prose rimée et rythmée ", *Le médiéviste et l'ordinateur*, 33 (printemps 1996), p. 41-48 (consultable sur Internet, sur le site Ménestrel).

[4] - La communication de M. Paulmier et M.-J. Gasse était optimistement intitulée " Une fausse piste provisoire : le futur antérieur dans les chartes médio-latines ". La piste s'est révélée définitivement fausse aux yeux des auteurs, qui affirment n'avoir découvert que des banalités ne justifiant pas une publication. Nous avons respecté leur souhait.

indications et définitions préalables ; un glossaire regroupe en outre, à la fin du volume, les termes techniques employés par les différents auteurs.

Pourquoi compter les mots ?

Dans un ouvrage paru il y a une dizaine d'années, Pierre Salat citait en exergue cette boutade d'Alexandre Vialatte : " La statistique est une science étonnante. Elle donne des certitudes chiffrées. Elle a prouvé que dans huit cas sur dix les boulangers sont des hommes qui fabriquent du pain. Ce qui confirme le pressentiment qu'on avait déjà de cette affaire, mais sans preuve scientifique et par pure intuition. Et voilà ce qu'il y a de beau dans la statistique : ce qu'on savait bêtement avant elle, on le sait ensuite scientifiquement. "[5] Il y a de cela, en effet, mais dans le cas où l'intuition et l'empirisme sont inopérants ou insuffisants, tout particulièrement dans les traitements de corpus textuels très grands, seule la statistique peut aider à y voir clair. La statistique textuelle est une porte d'entrée dans l'analyse linguistique et stylistique, elle permet de comparer des textes, de dégager les mots ou les groupements de mots favoris d'un auteur, etc… Les philologues et les historiens peuvent lui demander d'étayer des attributions de textes restés anonymes, finalité discutée dans plusieurs des contributions de ce volume, avec une forte insistance sur les difficultés et les dangers théoriques d'une telle utilisation, car un même auteur peut varier son lexique et son style d'une œuvre à l'autre, et, inversement, un auteur peut en imiter, en citer, en pasticher ou en parodier, un autre. La statistique textuelle établira alors de façon très fiable une ressemblance ou une différence entre les textes analysés, mais elle ne permettra pas de se prononcer sur l'identité des auteurs[6]. En revanche elle viendra corroborer des hypothèses ou des certitudes acquises sur d'autres critères[7]. Faite sur un large corpus, la statistique textuelle permet aussi de déceler des mouvements diachroniques, de décrire l'évolution d'une langue[8], de repérer des faits marquants ou étonnants ; la présence d'" intrus " dans un corpus diplomatique peut aider à déceler des actes faux, comme le montre la contribution de Nicholas Brousseau.

Spécificité de l'enquête statistique portant sur le latin

A la différence du français moderne[9], le latin est une langue à flexion, ou, si l'on préfère, à déclinaison. Cette particularité du latin a des incidences sur le traitement statistique des textes : faut-il compter les formes ou les lemmes (voir le glossaire) ? Si l'on décide de compter les lemmes, il faut *lemmatiser* les textes, c'est-à-dire ramener toutes les formes (par exemple *reges*, *regum*, *regibus*) au lemme ou vocable correspondant (*rex*). L'article de Sylvie Mellet traite des avantages et inconvénients des deux méthodes ; celui de Michel Dubrocard et Xuan Luong propose des études statistiques sur des formes. Cette dernière méthode est évidemment plus rapide, et elle autorise l'utilisation de logiciels conçus pour le français moderne, comme *Hyperbase*[10]. Elle fonctionne excellemment dans certains cas, réserve même des surprises, mais s'avère insuffisante dans d'autres[11].

---

[5] - P. Salat, " *Verborum ratio* ". Exemples d'études statistiques portant sur le vocabulaire latin, Clermont-Ferrand, 1991 (Faculté des lettres et Sciences humaines de l'Université Blaise-Pascal, fasc. 33).

[6] - Sur ce point, développé dans l'article d'E Evrard, il faut lire aussi F. Dolbeau, " Critique d'attribution, critique d'authenticité. Réflexions préliminaires ", *Filologia mediolatina*, VI-VII (1999-2000), p. 33-61.

[7] - Voir F. Dolbeau, " Recherches sur les œuvres littéraires du pape Gélase II. A- Une Vie inédite de Grégoire de Naziance (BHL 3668d) attribuable à Jean de Gaète ", *Analecta Bollandiana*, 107, p. 65-127, spéc. p. 95-97, qui confirme par une étude statistique du cursus (c'est-à-dire du rythme des fins de phrases) son attribution à Jean de Gaète d'une Vie de Grégoire de Naziance.

[8] - Voir par exemple B.-M. Tock, " Les mutations du vocabulaire latin des chartes a XI^{e} siècle ", dans *Pratiques de l'écrit documentaire au XI^{e} siècle*, Bibliothèque de l'Ecole des Chartes (B.E.C.) 155 (1997), p. 119-148, et Id., " Le latin médiéval et l'ordinateur ", dans M. Goullet et M. Parisse éd., *Les Historiens et le latin médiéval*, Paris, 2001, p. 55-65.

[9] - Il reste en français moderne quelques traces de déclinaison, par exemple dans les pronoms relatifs ou personnels : *je*, *moi*, *me* sont des formes différentes d'un même vocable, de même que *qui* et *que*.

[10] - Conçu par Etienne Brunet, et diffusé par le laboratoire " Bases, corpus et langage " de Nice. Voir S. Mellet, " Les tragédies de Sénèque vues à travers Hyperbase ", dans *Mots chiffrés et déchiffrés*, Mélanges offerts à Etienne Brunet, éd. S. Mellet et M. Guillaume, Paris, Champion, 1998, p. 255-272.

[11] - Voir S. Mellet, " Les atouts de la lemmatisation ", dans *Bases de données linguistiques : conceptions, réalisations, exploitations*. Actes du Colloque international de Corte (11-14 octobre 1995), éd. G. Moracchini, p. 309-316.

Les méthodes quantitatives ne se limitent pas au traitement lexical. Elles permettent aussi d'analyser des faits grammaticaux, comme le montre l'article de Sylvie Mellet. Cette opération est un peu plus longue que l'autre, car il convient de faire un encodage préalable, c'est-à-dire d'analyser chaque mot dans la phrase, et d'affecter à ce mot un code qui lui sert en quelque sorte de carte d'identité : sous un code alphanumérique (chiffres et lettres), on récapitule sa nature (nom, adjectif), sa fonction (sujet, complément), sa place dans le texte, etc… les critères pouvant être multipliés *ad libitum*. Au sortir d'une telle opération on peut réaliser des statistiques fines sur une catégorie grammaticale précise. L'encodage se fait de façon semi-automatisée : soit la forme *venere* ; l'ordinateur proposera deux analyses possibles (abl. de *venus, eris*, et 3e pers. pl. parf. de *venio*) ; c'est à l'utilisateur qu'il revient de " désambiguïser ", c'est-à-dire de lever l'homonymie en cochant l'analyse adéquate. On voit que pour les textes longs l'opération est très lourde, et en tout état de cause elle nécessite un logiciel spécial. Le Lasla de Liège effectue ce genre de traitement pour les chercheurs qui le souhaitent ; le paiement se fait au nombre de mots traités[12].

Lexicographie et sémantique

Une chose est d'isoler et compter les mots, autre chose de comprendre leur sens. La sémantique est précisément la discipline qui traite du langage considéré du point de vue du sens, la sémantique historique s'intéressant à l'évolution du sens des mots à travers le temps[13]. Une analyse exemplaire est proposée ici par Anita Guerreau et Bruno Bon à partir du vocable *pietas*, dont les sens oscillent, *grosso modo*, entre ceux du français " pitié " et " piété ", sans les recouvrir tout à fait. Faut-il pour autant renoncer à traduire ? Nous ne le pensons pas, et ce pour de multiples raisons. Tout d'abord si les vocables français " pitié " et " piété " - et leurs équivalents dans les autres langues romanes - se sont inscrits dans une filiation directe avec le latin, c'est forcément en vertu d'une continuité sémantique dont il nous appartient de retrouver le fil. D'autre part la question de la légitimité des traductions est un faux problème, vieux comme celui du rapport des mots aux choses qu'ils désignent (voir le *Cratyle*) : la traduction française de *pietas* ou de tout autre mot latin antique ou médiéval n'est pas plus illégitime que celle d'un mot de l'argot de Harlem ou du Bronx dans le sous-titrage d'un film américain. Le monde médiéval, dont nous avons en partie hérité, ne nous est pas plus étrange(r) qu'un ghetto de l'Amérique actuelle. Si traduire est toujours une approximation, c'est parce que dans toutes les langues le mot échoue à dire fidèlement la chose (voir Mallarmé). Il faut nous en accommoder, ou ne plus communiquer. L'article de Michel Parisse rend précisément compte d'un phénomène de traduction dans les chartes : la formule *quod vulgo dicitur*, associant à un mot latin un mot vernaculaire, est un témoignage de la diglossie médiévale : le latin est réservé à l'usage savant écrit, le français à l'usage courant oral, l'introduction du français dans un texte écrit entraînant d'ailleurs parfois une latinisation de sa finale.

De ce très dense échange de vues on retiendra la fécondité de ces méthodes, pourvu qu'elles soient bien employées et que nos questions soient bien posées[14]. Toute personne intéressée peut prendre contact avec l'un des deux laboratoires représentés ici (Lasla ou " Bases, corpus et langage "), qui les guidera en vertu des objectifs qu'il s'est fixés. En effet bases de données et statistiques ne peuvent être que des outils, dont la valeur est nulle si le médiéviste ne leur soumet pas les bonnes questions ou s'il interprète mal leurs résultats. Ce numéro un peu aride n'a donc pas pour but de " relooker " la médiévistique selon des critères " high tech ", mais de progresser dans l'évidence que nous construisons l'histoire en très grande partie sur les mots que nous a transmis le passé.

---

[12] - Lasla (Laboratoire d'Analyse Statistique des Langues Anciennes), Université de Liège, 1b, Quai Roosevelt, B-4000 Liège.

[13] - Sur cette discipline on peut lire A. Guerreau, *L'avenir d'un passé incertain*, Paris, 2001, p. ****, qui propose une excellente vision de ce qu'est une sémantique historique bien conçue, mais qui sous-estime certainement la prise de conscience et les connaissances de bon nombre de ses collègues. En outre à sa bibliographie presque exclusivement allemande il faudrait ajouter les très nombreux travaux des Anglo-saxons et des Français, que nous ne pouvons énumérer ici.

[14] - Sur le danger de certains présupposés dans les enquêtes statistiques on peut lire A. Guerreau, " A propos d'une liste de fréquence des dénominations professionnelles dans la France du XIXe siècle ", *Annales ESC*, 4 juillet-août 1993, p. 979-986.

Petite bibliographie

- Statistiques textuelles :
- L. Lebart et A. Salem, *Statistique textuelle*, Paris, Dunod, 1994.
- Ch. Muller, *Initiation aux méthodes de la statistique linguistique*, Paris, Champion, 1992 (réimpression de l'édition de 1973).
- Id., *Principes et méthodes de la statistique lexicale*, Paris, Champion, 1992 (réimpression de l'édition de 1977).

- Méthodes quantitatives appliquées au latin :
- Michel Dubrocard, " *César* dans César ", dans *Travaux du cercle linguistique de Nice*, 16, 1994.
- Id., " Cooccurrences significatives et dimensions du contexte *César* dans César (suite) ", dans *Mots chiffrés et déchiffrés*. Mélanges offerts à Etienne Brunet, éd. S. Mellet et M. Guillaume, Paris, Champion, 1998, p. 67-81.
- Etienne Evrard, " Etude métrique du *Carmen de Sancto Landberto* ", dans *Mots chiffrés et déchiffrés*, p. 101-112.
- Id., " Pour un inventaire raisonné de la syntaxe latine ", dans *Serta Leodiensia secunda*. mélanges publiés par les Classiques de Liège à l'occasion du 175e anniversaire de l'Université, Liège, 1992, p. 173-190.
- Etienne Evrard et Sylvie Mellet, " Les méthodes quantitatives en langues anciennes ", *Lalies* 18 (1998), p. 111-155.
- X. Luong, " Le consensus en analyse arborée ", dans *Mots chiffrés et déchiffrés*, p. 187-197.
- Sylvie Mellet, " Les atouts de la lemmatisation ", dans *Bases de données linguistiques : conceptions, réalisations, exploitations*. Actes du Colloque international de Corte (11-14 octobre 1995), éd. G. Moracchini, p. 309-316.
-Ead., les tragédies de Sénèque vues à travers Hyperbase, dans *Mots chiffrés et déchiffrés*, p. 255-272.
- Benoît-Michel Tock, " Les mutations du vocabulaire latin des chartes au XIe siècle ", *Bibliothèque de l'Ecole des chartes*, t. 155 (1997), p. 119-148.
- Philippe Verkerk et Anne-Marie Turcan Verkerk, " Un programme informatique pour l'étude de la prose rimée et rythmée ", *Le médiéviste et l'ordinateur*, 33 (printemps 1996), p. 41-48.

Critique d'attribution :
Le n° VI-VII de *Filologia mediolatina* (SISMEL, 1999-2000) est entièrement consacré aux problèmes d'attribution. On signalera surtout, pour ce qui nous occupe ici :
- Giovanni Orlandi, " Metrica e statistica linguistica come strumenti nel metodo attributivo ", p. 9-31 ; [voir aussi Id., " Le statistiche sulle clausole della prosa. Problemi e proposte ", *Filologia mediolatina* V (1998).]
- F. Dolbeau, " Critique d'attribution, critique d'authenticité. Réflexions préliminaires ", p. 33-61.

Lexique médio-latin
- La revue Alma (Archivum Latinitatis Medii Aevi) ou Bulletin du Du Cange est entièrement consacrée aux études lexicales.
- M. Goullet et M. Parisse éd., *Les Historiens et le latin médiéval*, Paris, 2001 (Actes du colloque tenu à la Sorbonne en septembre 1999).

Glossaire

Analyse factorielle : l'analyse factorielle permet de mesurer les distances et les proximités (autrement dit les **connexions***), entre les textes en extrayant les principaux éléments qui les rapprochent ou les séparent. Des logiciels (du type Hyperbase) transforment automatiquement les données numériques en une représentation graphique sur laquelle les textes qui se ressemblent du point de vue des critères adoptés figurent dans des zones voisines.

Analyse morphosyntaxique : l'analyse morphologique porte sur la forme du mot (cas, genre et nombre d'un susbstantif, par ex.), tandis que l'analyse syntaxique porte sur ses rapports avec les autres mots de la phrase (fonction pour un substantif, modes verbaux dont sont suivies les conjonctions, etc…). Le plus souvent la statistique regroupe les deux analyses sous le terme " morphosyntaxique ".

Bruit, bruité : le terme **bruit** est employé en statistique de façon imagée avec un sens voisin de celui qu'il a en acoustique, où il désigne un phénomène aléatoire gênant qui vient parasiter un signal utile. Une expérience est dite **bruitée** si elle est faussée par des phénomènes indésirables et mal contrôlés.

Chaîne de caractères : voir **Mot graphique**.

Code alphanumérique : voir **Encodage**.

Désambiguïsation : lorsqu'on demande à l'ordinateur d'analyser un texte latin, il n'est pas capable de résoudre seul certaines difficultés suscitées par des phénomènes d'homonymie (*actum* est-il une forme du substantif *actus, us*, ou du verbe *ago* ?). L'utilisateur doit donc procéder à une **levée d'homonymie** ou **désambiguïsation**, en choisissant l'analyse convenable parmi celles que propose la machine.

Connexion entre les textes : l'évaluation de la **connexion** entre deux textes peut s'énoncer en termes de **proximité** (ce qu'ils ont en commun) ou de **distance** (ce qui les différencie). Si la statistique porte sur le lexique, il y a proximité entre deux textes quand une grande partie de leur vocabulaire est commun, et distance quand leur vocabulaire est très différent ; le rapport entre les deux mesure la connexion de ces textes.

Encodage : l'encodage, ou **étiquetage**, consiste à affecter d'un certain nombre de chiffres et de lettres les formes d'un texte qu'on veut analyser (d'où le terme **alphanumérique**). Ce **code** ou **étiquette** indique, par exemple, la place du mot dans le texte, sa nature grammaticale, sa fonction dans la phrase si c'est un substantif, son temps, sa voix et sa personne si c'est un verbe, etc, les codes retenus variant selon le type de résultats recherchés.

Espérance mathématique : valeur théorique* que laisse attendre le calcul des probabilités.

Etiquetage : v. **Encodage**.

Forme : v. **Lemmatisation**.

Graphème : plus petite unité graphique (synonymes courants : " caractère graphique " ou " lettre " ; mais ce peut être aussi un signe comme l'apostrophe, ou un signe de ponctuation).

χ^2 (khi deux ou khi carré) : le test du χ^2 ou test de Pearson sert à apprécier l'écart constaté entre une observation réelle et un modèle théorique, donc à se prononcer sur le caractère statistiquement significatif d'un résultat. Ainsi au jeu de pile ou face, le modèle théorique, si la pièce est homogène dans sa structure, et les joueurs honnêtes, laisse prévoir un nombre égal de piles et de faces, si le nombre des lancers est assez grand. Supposons que sur 100 lancers, on ait obtenu 40 piles

et 60 faces au lieu des 50 attendus - effectif théorique des piles et des faces dans l'hypothèse d'une distribution aléatoire -, l'écart pour les piles est de -10 pour les piles, de +10 pour les faces. Le χ^2 s'obtient en élevant au carré ces écarts, et en les divisant par l'effectif théorique. Nous aurons ici pour chacun des deux éléments ($10^{10}/50 = 100/50 = 2$. Le total des χ^2 partiels est égal à 2+2 soit 4. Il y a donc un peu moins de 5 chances sur 100 pour que le hasard permette d'expliquer les écarts observés. De la même façon, si dans un corpus constitué par la réunion de deux textes, A et B, de dimensions égales, une forme est employée 100 fois, et si elle figure 40 fois dans le texte A et 60 fois dans le texte B, il y a un peu moins de 5 chances sur 100 pour que l'excédent et le déficit observés soient l'effet du hasard. En affirmant que ces écarts ont une autre explication (auteurs différents, genres différents, sujets différents, etc.), on a un peu plus de 95 chances sur 100 de ne pas se tromper. Une table des valeurs du χ^2, que l'on trouve dans les manuels de statistiques, permet de constater que la probabilité associée est un peu inférieure à 5%. Des programmes informatiques de statistiques textuelles calculent automatiquement le χ^2.

Lemmatisation : en latin ce que le langage courant appelle un même mot (par exemple *rex*, " roi "), prend différentes **formes** selon sa fonction dans la phrase (par exemple, en fonction de complément d'objet direct, *regem* (singulier) et *reges* (pluriel). On dira donc que *reges*, comme *rex*, *regem*, *regibus*, etc… sont des **formes** différentes d'un même **vocable** ou **lemme** que l'on regroupe sous le nominatif *rex*. Dans un texte latin il y a donc plus de formes que de vocables ou de lemmes. On classe les vocables ou lemmes sous le nominatif singulier pour les substantifs (*rex*), sous le nominatif masculin singulier pour les adjectifs (*magnus*), et sous la première personne du singulier de l'indicatif présent actif (parfois sous l'infinitif présent actif) pour les verbes (*amo*, parfois *amare*).

Lemme : v. **Lemmatisation**

Levée d'homonymie : v. **Désambiguïsation**

Mot graphique : lorsqu'on dit qu'un texte latin contient 2500 mots, il s'agit du nombre d'unités graphiques séparées par un blanc (*filius regis Italie* = 3 mots graphiques) : en statistique, pour éviter la confusion avec lemme ou vocable, on parle souvent de **mots graphiques**. Si l'unité retenue n'est pas le mot, on parlera de **chaîne de caractères** (terminaison d'un verbe ou d'un nom, préfixe, suffixe, etc…)

Numériser : transformer des données en une suite de valeurs numériques, qui permet leur traitement informatique.

Quadrant : chacune des quatre portions du plan délimitées par un système de coordonnées rectangulaires.

Théorique (effectif, modèle, vocabulaire) : en statistiques est dit **théorique** ce qui s'obtient par le calcul des probabilités. Le modèle théorique s'oppose au modèle **effectif** ou **réel**, observé lors d'une expérience. La différence entre le modèle théorique (= attendu) et le modèle effectif (= observé) permet de déterminer les caractères significatifs d'un texte.